BIOMETRICS 53, 1440–1457
December 1997

Design Aspects of Calibration Studies in Nutrition, with Analysis of Missing Data in Linear Measurement Error Models

Raymond J. Carroll,[1] Laurence Freedman,[2] and David Pee[3]

[1]Department of Statistics, Texas A&M University,
College Station, Texas 77843-3143, U.S.A.
[2]Biometry Branch, DCPC, National Cancer Institute,
Executive Plaza North, Room 344, MSC 7354,
Bethesda, Maryland 20892-7354, U.S.A.
[3]Information Management Services, Inc., 6120 Executive Boulevard,
Rockville, Maryland 20852, U.S.A.

SUMMARY

Motivated by an example in nutritional epidemiology, we investigate some design and analysis aspects of linear measurement error models with missing surrogate data. The specific problem investigated consists of an initial large sample in which the response (a food frequency questionnaire, FFQ) is observed and then a smaller calibration study in which replicates of the error prone predictor are observed (food records or recalls, FR). The difference between our analysis and most of the measurement error model literature is that, in our study, the selection into the calibration study can depend on the value of the response. Rationale for this type of design is given. Two major problems are investigated. In the design of a calibration study, one has the option of larger sample sizes and fewer replicates or smaller sample sizes and more replicates. Somewhat surprisingly, neither strategy is uniformly preferable in cases of practical interest. The answers depend on the instrument used (recalls or records) and the parameters of interest. The second problem investigated is one of analysis. In the usual linear model with no missing data, method of moments estimates and normal-theory maximum likelihood estimates are approximately equivalent, with the former method in most use because it can be calculated easily and explicitly. Both estimates are valid without any distributional assumptions. In contrast, in the missing data problem under consideration, only the moments estimate is distribution-free, but the maximum likelihood estimate has at least 50% greater precision in practical situations when normality obtains. Implications for the design of nutritional calibration studies are discussed.

1. Introduction

1.1 *Overview*

The assessment and quantification of an individual's usual diet is a difficult exercise, but one that is fundamental to discovering relationships between diet and disease and to monitoring dietary behavior among individuals and populations. Various dietary assessment instruments have been devised, of which three main types are most commonly used in contemporary nutritional research. The one that is most convenient and inexpensive to use is the food frequency questionnaire (FFQ), which is the instrument of choice in large nutritional epidemiology studies. FFQs are structured lists of foods commonly consumed by the target population arranged in food groups (such as meats, fruits and vegetables, breakfast cereals, etc.). The respondent is required to go down the list and indicate the frequency (never, once a month, once a week, etc.) with which each item has been generally consumed over a recent long-term period, say 12 months. Some FFQs include specification

Key words: Errors-in-variables; Estimating equations; Linear regression; Maximum likelihood; Measurement error; Method of moments; Missing data; Model robustness; Nutrition; Sampling designs; Semiparametrics; Stratified sampling; Weighting.

of the average portion size consumed. While dietary intake levels reported from FFQs are correlated with true usual intake, they are thought to involve a systematic bias (i.e., under- or overreporting at the level of the individual). This is due partly to the difficulty of recalling diet over a long-term period, partly to the difficulty of specifying portion size, and partly to the difficulties of converting the limited information on the list into exact amounts of nutrients.

The other two instruments that are commonly used are the 24-hour food recall and the multiple-day food record (FR). Twenty-four-hour food recalls are obtained by a trained interviewer questioning a respondent on the food that was consumed the previous day, including full details of brand names of products and portion sizes. Food records are self-completed records of food consumed over a multiple-day period, with the respondent ideally recording the details of foods as they are consumed. Each of these is more work-intensive and more costly but is thought to involve less bias than an FFQ because the meals taken are either very recent or current and the details of each food item can be specified more precisely. However, the large daily variation in a Western diet makes a single FR an imprecise measure of true usual intake.

1.2 *Calibration Studies and Their Aims*

Despite the problem that FRs are not completely unbiased and may involve some underreporting, their generally accepted superiority over FFQs makes them the current practical gold standard for dietary assessment. Thus, for proper interpretation of epidemiologic studies that use FFQs as the basic dietary instrument, one needs to know the relationship between reported intakes from the FFQ and true usual intake, defined by the average intakes reported over a very long series of FRs. Such a relationship is ascertained through a substudy, commonly called a calibration (or validation) study.

The design of calibration studies has only recently attracted the interest of biostatisticians. This interest arises from the growing awareness of the problem of error in the measurement of exposures and its effect on estimation and power in epidemiologic studies. Calibration studies can provide valuable information on the nature and magnitude of the error using a given measurement method and are therefore important for the proper design and interpretation of epidemiologic studies using that method. Currently, the epidemiologic area that is most actively engaged in the conduct of calibration studies is nutrition, probably because of the profound problems in measuring dietary intake. Most of the calibration studies that have been conducted in this area have been associated with a larger epidemiologic study using the same measurement instrument. In early calibration studies, it was assumed that one should attempt to measure the usual dietary intake of an individual as accurately as possible to act as a gold-standard comparison with the more approximate instrument to be used in the larger study (e.g., Willett, Sampson, and Stampfer, 1985). However, this wisdom was challenged when statisticians considered how such data would be used to aid interpretation of the main study. First, Carroll et al. (1984) and Rosner, Willett, and Spiegelman (1989) demonstrated that a valid measurement error adjustment of the relative risk estimates from the main study can be made even when the calibration study does not include a very accurate measure of usual dietary intake. It would be sufficient to include a measure of intake that is simply unbiased. This meant that it was not imperative to use many repeat measurements of dietary intake in order to greatly increase the precision of the measure. Then, Kaaks, Riboli, and van Stavern (1995) and Stram, Longnecker, and Shames (1995) considered the variance of the estimated relative risks adjusted for measurement error using such data from a calibration study. They discovered that, if one has a choice between increasing the number of repeat measurements per individual or increasing the number of individuals, then under assumptions of equal cost, the variance is minimized by maximizing the number of individuals in the calibration study. Stram et al. (1995) also investigated the optimal strategy when costs are not equal.

The primary aim of a calibration study may not be exactly the same in each case. In this paper, we consider four possibilities.

(a) The aim that has been most frequently described is to use information from the calibration study to adjust the relative risks estimated from the main epidemiologic study for the measurement error associated with use of the FFQ (Kaaks et al., 1995). It is well-known that measurement error biases the estimated relative risks, and this motivates the need for adjustment.

(b) Another possible aim is to estimate the sample size required in the main study. The required sample size depends heavily on the degree of measurement error associated with the FFQ (Freedman, Schatzkin, and Wax, 1990), so there is good reason to check on this before proceeding with the main epidemiologic study. In this case, it is important that the calibration study be conducted and evaluated before the main study proceeds.

(c) A third possible aim is to estimate the correlation between FFQ intake and true usual intake. This could be of crucial interest if the FFQ has been modified extensively from previous versions or is to be used in a new population from which little previous data have been obtained. Very low correlations might persuade the investigators to postpone the main study, pending improvements in the design of the FFQ or in the way it is presented to study participants.

(d) A fourth possible aim is to estimate the slope of the regression of FFQ intake on the usual intake. This parameter is of importance in assessing the patterns of bias that might exist with use of the FFQ.

1.3 *The American Association for Retired People (AARP) Study*

The National Cancer Institute and American Association for Retired People (AARP) are collaborating in conducting a large prospective nutritional epidemiologic study, in which members of the AARP will report information on their dietary habits and will be followed to ascertain new diagnoses of cancer. The motivation for the study was, first, the degree of disagreement and controversy over the results of previous epidemiologic studies of diet and cancer, particularly breast cancer (Prentice et al., 1988); second, the limited range of intakes of major macronutrients, such as fats, in previously studied cohorts (Hebert and Miller, 1988); and third, the need for large numbers of cancer cases to occur during follow-up for the detection of small but important observed relative risks. The last point is emphasized by noting that a true relative risk of 2.0 can typically be reduced by dietary measurement error to an observed relative risk of 1.25 (Freudenheim and Marshall, 1988).

The design of the main AARP study involves a two-stage sampling. First, a large number of randomly selected members are sent an FFQ to complete. Second, a group of the respondents are selected using stratified random sampling on the basis of their reported intake on a selected macronutrient of interest, e.g., fat. The stratified sampling ensures that subjects with extremely high or low reported intakes have a high probability of being selected while those with reported intakes closer to the average would have a lower probability of selection. Using percent calories from fat as the intake measure and five strata of intake (<25%, 25–32.5%, 32.5–40%, 40–47.5%, >47.5%), the initially estimated required sample size for the cohort is 350,000 (Freedman, Schatzkin, and Wax, 1991b).

Besides the main study, there is also a calibration study. This is an important part of the project, particularly because there is not wide experience with the results of mailing dietary questionnaires. There are three main aims of the calibration study: to check on the correlation between the reported FFQ intake and true usual intake to see if the mailed responses to the questionnaire have adequate validity, to check on the estimated sample size required in the main study, and to correct relative risk estimates from the main study.

Participants in the calibration study are to be randomly selected from respondents in the first stage of the study. Two options for this random selection suggest themselves. First, one might simply take a simple random sample of the respondents and ask them to complete one or more FFQs and one or more FRs or, alternatively, one might design the calibration study to parallel the main study and preferentially select those individuals who report extreme intakes on their FFQs.

1.4 *Questions Posed in This Paper*

In this paper, we analyze two aspects of the design of the calibration study. First, do we gain or lose efficiency, especially with regard to questions (a)–(d) in Section 1.2, by taking a stratified random sample? Second, is it better to obtain many FRs from a moderate number of individuals or only a small number of FRs on a larger number of individuals? Many researchers take the first option so that they can characterize usual intake for each individual as accurately as possible. However, Kaaks et al. (1995) argues that, to achieve optimal adjustment of relative risks, it is better to take only one FR per person and thus maximize the number of persons in the calibration study. Also, Rosner and Willett (1988) show that, for estimating the correlation between an FFQ and usual intake, the optimal design depends on the amount of error in the FRs. We investigate these design options to see whether the optimal strategy depends on the different possible aims of a calibration study, as described in Section 1.2.

The second question concerns analysis. Because we are unable to measure usual intake precisely, we are necessarily in the realm of errors-in-variables analysis (see models (1) and (2) below). It is our experience that, for simple random sampling, the method of moments and normal-theory maximum likelihood estimates in linear measurement error models have approximately the same efficiencies, and hence the former is useful because it has explicit formulas. We investigate these methods for the case when we sample from strata defined by the values of the response and, as described below, we show that, in this case, the two methods have surprisingly different behavior.

The paper is organized as follows. In Section 2, we discuss the linear measurement error model and place the AARP calibration study into this framework. In Section 3, we discuss estimation in the context of missing data. The main conceptual device is to place linear errors-in-variables estimation into the framework of unbiased estimating functions. Using results of Rotnitzky and Robins (1995), we also show how to obtain the asymptotically optimal estimator that makes no distributional assumption. Sections 4 and 5 contain numerical results. In Section 6, we discuss the implications of our results. Some technical details are collected into the Appendix.

2. Statistical Model for Calibration

As described previously, the AARP calibration study involves selecting a random sample from respondents to the first stage of the main study. At this first stage, for a large number M of individuals, nutrient intake is measured by an FFQ. In the calibration study, on a smaller number n of individuals, the FFQ nutrient intake is calibrated against usual intake by measuring nutrient intake with two or more food records or recalls (FR), possibly together with additional FFQs.

Our analysis is based on the general statistical calibration model of Freedman, Carroll, and Wax (1991a). They allow for the possibility that one or more FFQs are measured contemporaneously with FRs and hence that the errors are correlated. Here we will assume the simpler case that the FFQs and the FRs are measured sufficiently far apart that all errors are uncorrelated.

Consider persons randomly selected to participate in the calibration study. The individual reports diet using an FFQ on m_1 occasions ($m_1 \geq 1$) and using an FR on m_2 occasions ($m_2 \geq 2$). The model relating intake of some nutrient (e.g., percent calories from fat) reported on FFQs (denoted by Q) and intake reported on FRs (denoted by F) to long-term usual intake (denoted by T) is a standard linear errors-in-variables model, namely,

$$Q_j = \beta_0 + \beta_1 T + r + \epsilon_j; \qquad j = 1, \ldots, m_1; \tag{1}$$

$$F_j = T + U_j; \qquad j = 1, \ldots, m_2. \tag{2}$$

In model (1), r is called the equation error (Fuller, 1987). The terms ϵ_j represent the within-individual variation in FFQs, while the U_j are the within-individual variation in FRs.

For example, in the AARP calibration study, an FFQ will be obtained initially, and some months later an FR will be obtained, followed by a second FR obtained at least 1 month later. Then $m_1 = 1$ and $m_2 = 2$. If, subsequently, a second FFQ is obtained, then $m_1 = 2$.

Among these random variables, T has mean μ_t and variance σ_t^2, U_j has mean zero and variance σ_u^2, ϵ_j has mean zero and variance σ_ϵ^2, and r has mean zero and variance σ_r^2. Note the critical assumption that U_j has mean zero, i.e., that the FR provides an unbiased measurement of dietary intake. All random variables are uncorrelated, although the methods are easily extended to allow for correlation between the pairs of measurement errors ϵ_j and U_j corresponding to a questionnaire being given nearly coincidentally in time to a record or recall. The parameter σ_ϵ^2 cannot be estimated if $m_1 = 1$, i.e., if there are no replicated FFQs and, in this case, the remaining parameters are estimated by setting $\sigma_\epsilon^2 = 0$; the estimate of σ_r^2 then incorporates the contribution of σ_ϵ^2. If $m_2 = 1$, then the measurement error variable σ_u^2 cannot be estimated, and it is well known that β_1 cannot then be estimated (Fuller, 1987).

In what follows, it is convenient to reparameterize the problem in terms of means, variances, and covariances. Make the definitions $\theta_1 = \mathrm{E}(Q)$, $\theta_2 = \mathrm{E}(T)$, $\theta_3 = \mathrm{var}(Q)$, $\theta_4 = \mathrm{cov}(Q, F)$, $\theta_5 = \mathrm{var}(F)$, $\theta_6 = \mathrm{cov}(F_1, F_2)$, and $\theta_7 = \mathrm{cov}(Q_1, Q_2)$. Let $\boldsymbol{\Theta} = (\theta_1, \theta_2, \ldots, \theta_7)^t$ (where superscript t indicates transpose), and let $\mathbf{e}_k$ be the vector of k 1s. All the model parameters can be obtained from $\boldsymbol{\Theta}$, specifically,

$$\beta_1 = \theta_4/\theta_6; \qquad \mu_t = \theta_2; \qquad \beta_0 = \theta_1 - \theta_2\theta_4/\theta_6; \qquad \sigma_t^2 = \theta_6;$$
$$\sigma_u^2 = \theta_5 - \theta_6; \qquad \sigma_r^2 = \theta_7 - \theta_4^2/\theta_6; \qquad \sigma_\epsilon^2 = \theta_3 - \theta_7.$$

If $m_1 = 1$, then θ_7 cannot be estimated and (using the convention that $\sigma_\epsilon^2 = 0$) $\sigma_r^2 = \theta_3 - \theta_4^2/\theta_6$.

The possible observed data are summarized as $\mathbf{Z} = (Q_1, \ldots, Q_{m_1}, F_1, \ldots, F_{m_2})^t$, which has mean $(\theta_1 e_{m_1}^t, \theta_2 e_{m_2}^t)^t$ and covariance matrix

$$\boldsymbol{\Sigma}(\boldsymbol{\Theta}) = \begin{bmatrix} \theta_3 \mathbf{I}m_1 + \theta_7(\mathbf{e}_{m_1}\mathbf{e}_{m_1}^t - \mathbf{I}m_1) & \theta_4 \mathbf{e}_{m_1} e_{m_2}^t \\ \theta_4 \mathbf{e}_{m_2} e_{m_1}^t & \theta_5 \mathbf{I}m_2 + \theta_6(\mathbf{e}_{m_2}\mathbf{e}_{m_2}^t - \mathbf{I}m_2) \end{bmatrix}. \tag{3}$$

In the above statistical model, equation (2) indicates an assumption that the food record or recall is an unbiased estimate of the individual's usual intake and that the mean of a sufficient number of repeated FR values is arbitrarily close to the true usual intake. Recent research has thrown into question this assumption. Plummer and Clayton (1993) showed that protein intake was underestimated by both food records and recalls compared to values calculated from 24-hour urinary nitrogen excretion that are thought to be close to true intake, while Heitmann and Lissner (1995) showed that food records underestimated total energy intake compared to more exact values calculated by the double-labeled water method. However, it is still unclear whether nutrient densities, in particular percent calories from fat, are also underestimated or whether there is a general underreporting phenomenon that applies equally to all nutrients. If some nutrients, such as fat, are underreported more than others, then our model would indeed be misspecified. The implications of such possible misspecification for the adjustment of disease relative risks obtained from nutritional epidemiologic studies is an area of active development (see Prentice, 1996; Carroll et al., 1998). However, if the underreporting is general, i.e., applies as equally to fat as to other nutrients, then our model is a reasonable one, and we proceed on that basis for this paper.

3. The Two-Stage Study as a Missing Data Problem

3.1 *Introduction*

Because the AARP main study will preferentially select individuals who report more extreme levels of dietary intake, we may wish the calibration study to have similar composition. We therefore consider a calibration design where sampling is done in two stages. At the first stage, we observe the FFQs Q_{i1} for M individuals, $i = 1, \ldots, M$. Then at the second stage, the calibration study, with probability $\pi(Q_{i1})$, we observe the m_2 FRs $(F_{i1}, \ldots, F_{im_2})$ and the remaining $m_1 - 1$ FFQs $(Q_{i2}, \ldots, Q_{im_1})$. If an individual is selected into the calibration study, we set $\Delta_i = 1$, and otherwise we set $\Delta_i = 0$. The sampling weights are $w_i = 1/\pi(Q_{i1})$, the inverses of the probabilities of selection. In typical applications, the size of the calibration study is fixed, say to n observations, so that the Δ's are correlated.

This formulation allows for simple random sampling by setting $\pi(Q)$ to be a constant, i.e., independent of the report from the first FFQ. The classical linear measurement error model assumes complete sampling so that all individuals participate in the calibration study, and hence $\Delta_i = \pi(Q_{i1}) = 1$.

It is important to observe that this formulation is that of a missing data problem, wherein the FRs and the supplementary FFQs are missing for many individuals. As a result of the design, the data are missing at random, i.e., missingness depends only on the value of the first FFQ and not on the unobserved FFQs or FRs.

The purpose of this section is to discuss various estimation strategies. In Section 3.2, we discuss two basic estimating functions for complete data, one based on the method of moments and one based on maximum likelihood estimation. Section 3.3 describes adaptations of these estimating functions that allow for the missing data pattern of the AARP study. Section 3.4 gives some explicit details of the AARP study, which form the basis of all our later calculations. In practice, the sampling probabilities and hence the sampling weights are unknown and must be estimated (see also Section 3.4).

3.2 *Method of Moments and Model Robustness*

The typical measurement error model formulation has no missing data. The problem then is the classical linear measurement error model covered so admirably by Fuller (1987). With no missing data, there are two types of estimates in common use:

(a) Method of moments estimators expressed in terms of the model parameters $\mu_x, \sigma_x^2, \ldots$ and thus indirectly in terms of $\boldsymbol{\Theta}$. This is effectively the method used by Fuller (1987, pp. 106–108), and we will take it to be the default measurement error analysis.

(b) Maximum likelihood estimators assuming that all random variables are normally distributed and expressed in terms of the model parameters $\boldsymbol{\Theta}$.

With no missing data, these estimators can be expressed in terms of solutions to unbiased estimating equations. These methods solve equations of the form

$$0 = \sum_i \boldsymbol{\Psi}\left(\mathbf{Z}_i, \boldsymbol{\Theta}\right). \tag{4}$$

In what follows, i refers to the individual, Q_{ij} is the jth FFQ for the ith individual, $\bar{Q}_{i\cdot}$ is the within-individual mean, and similarly for F_{ij} and $\bar{F}_{i\cdot}$. Also, m_1 is the number of FFQs for each individual and m_2 is the number of FRs. We use the term $I(m_1 > 1)$ to be the indicator that $m_1 > 1$.

With no missing data, the estimating function for the method of moments is given by

$$\boldsymbol{\Psi}_{\text{mom}}(\mathbf{Z}_i, \boldsymbol{\Theta}) = \begin{bmatrix} \bar{F}_{i\cdot} - \mu_t \\ \Sigma_{j=1}^{m_2}(F_{ij} - \bar{F}_{i\cdot})^2 - (m_2 - 1)\sigma_u^2 \\ \begin{pmatrix} 1 & \bar{F}_{i\cdot} \\ \bar{F}_{i\cdot} & \bar{F}_{i\cdot}^2 - \sigma_u^2/m_2 \end{pmatrix}^{-1} \begin{pmatrix} \bar{Q}_{i\cdot} \\ \bar{Q}_{i\cdot}\bar{F}_{i\cdot} \end{pmatrix} - \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \\ (\bar{F}_{i\cdot} - \mu_t)^2 - \sigma_u^2/m_2 - \sigma_t^2 \\ (\bar{Q}_{i\cdot} - \beta_0 - \beta_1\bar{F}_{i\cdot})^2 - \sigma_\epsilon^2/m_1 - \beta_1^2\sigma_u^2/m_2 - \sigma_\tau^2 \\ \Sigma_{j=1}^{m_1}(Q_{ij} - \bar{Q}_{i\cdot})^2 - (m_1 - 1)\sigma_\epsilon^2 \end{bmatrix}.$$

When there is only one FFQ ($m_1 = 1$), we set $\sigma_\epsilon^2 = 0$ and remove the last component of this estimating function. The rows of $\boldsymbol{\Psi}_{\text{mom}}(\mathbf{Z}_i, \boldsymbol{\Theta})$ can be described as follows. Each row determines a parameter, i.e., row 1, mean of usual intake; row 2, within-individual measurement error variance in FRs; rows 3–4, intercept and slope of the regression model; row 5, variance of usual intake; row 6, equation error variance; row 7, within-individual error variance in FFQs.

The estimating function for the maximum likelihood estimator when there are no missing data is given in the Appendix, Section A.1.

3.3 *Model Robustness and Missing Data*

When there are no missing data, by solving (4), the method of moments and the maximum likelihood estimators are consistent and asymptotically normally distributed without restrictions as to the distributions of the random variables in the model. Of course, the asymptotic distributions depend on the underlying random variables, so that normal-theory information standard errors are valid only if all random variables are normally distributed. Otherwise, the simplest technique is to use sandwich standard errors: this is an old idea dating back at least to Huber (1967). There is also a sandwich-type theory of likelihood-ratio tests (see Huber, 1967; Kent, 1982).

With missing data, if the probability of selection into the calibration study is $\pi(Q_{i1})$, then we can construct consistent estimates as follows. For the method of moments, we use only the validation data and weight it inversely with the selection probabilities, thus solving

$$0 = \sum_{i=1}^{M} \Delta_i \boldsymbol{\Psi}_{\text{mom}}(\mathbf{Z}_i, \boldsymbol{\Theta})/\pi(Q_{i1}). \tag{5}$$

This approach is, of course, the well-known Horvitz–Thompson method (Horvitz and Thompson, 1952).

With missing data, the moments estimate obtained through solving (5) is still consistent and asymptotically normal even without assuming normality. The normal-theory maximum likelihood estimator, however, does not share this property. Here is a subtle point. This likelihood estimator, which ignores the missing data mechanism, may give inconsistent parameter estimates if the random variables are not all normally distributed. A brief explanation is given in the Appendix, Section A.2.

One can modify the likelihood estimator using the Horvitz–Thompson device to make it distribution-free, just as in (5). However, an asymptotically more efficient distribution-free estimator can be derived as follows.

The problem of estimating $\boldsymbol{\Theta}$ without making any distributional assumptions is semiparametric in the sense that parametric restrictions are made on the relationship of the means and variances of what we have called $\mathbf{Z}$, while the underlying distributions are nonparametric. Optimal estimation of $\boldsymbol{\Theta}$ in such a context has been discussed by Rotnitzky and Robins (1995). Here we discuss their methods and adapt them to our problem. We do not justify any of the theoretical claims made here, as they are either proved by Rotnitzky and Robins or simple consequences of their arguments.

Let $\mathbf{R}$ be the vector of all individual elements of $\mathbf{Z}$ and their cross products. For example, if $m_1 = 1$ and $m_2 = 2$, $\mathbf{Z} = (Q_1, F_1, F_2)^t$ and $\mathbf{R} = (Q_1, F_1, F_2, Q_1^2, Q_1F_1, Q_1F_2, F_1^2, F_2^2, F_1F_2)^t$. Let $\mathbf{g}(\boldsymbol{\Theta}) = \text{E}(\mathbf{R})$ and $\boldsymbol{\nu}(\boldsymbol{\Theta}) = \mathbf{R} - \mathbf{g}(\boldsymbol{\Theta})$. Of course, since we have specified the mean and covariance matrix of $\mathbf{Z}$, $\mathbf{g}(\boldsymbol{\Theta})$ can be computed without reference to underlying distributions. Define $\Delta = 1$ if an observation is selected into the calibration study and $\Delta = 0$ otherwise. Define

$$\underline{\mathbf{L}} = \left\{ \frac{\partial}{\partial \boldsymbol{\Theta}^t} \mathbf{g}(\boldsymbol{\Theta}) \right\}^t;$$

$$\boldsymbol{\chi}(\mathbf{Z}, \Delta, \boldsymbol{\Theta}) = \frac{\Delta \{\mathbf{R} - \mathrm{E}(\mathbf{R} \mid Q_1)\}}{\pi(Q_1)} + \mathrm{E}(\mathbf{R} \mid Q_1) - \mathbf{g}(\boldsymbol{\Theta});$$

$$\mathcal{T}_1(\boldsymbol{\Theta}) = \mathrm{cov}\{\boldsymbol{\chi}(\mathbf{Z}, \Delta, \boldsymbol{\Theta})\};$$

$$\mathcal{T}_2(\boldsymbol{\Theta}) = \left(\underline{\mathbf{L}} \mathcal{T}_1^{-1} \mathbf{L}^t\right)^{-1}.$$

Rotnitzky and Robins (1995) prove that the best that any semiparametric estimator of $\boldsymbol{\Theta}$ can achieve is an asymptotic covariance matrix of $M^{-1}\mathcal{T}_2(\boldsymbol{\Theta})$. They further show that the optimal estimating function for achieving this covariance matrix is to solve

$$0 = \sum_{i=1}^{M} \underline{\mathbf{L}} \{\mathcal{T}_1(\boldsymbol{\Theta})\}^{-1} \boldsymbol{\chi}(\mathbf{Z}_i, \Delta_i, \boldsymbol{\Theta}). \tag{6}$$

It is not entirely obvious that (6) is an unbiased estimating equation. To see this, one has to compute the expectation of $\boldsymbol{\chi}(\mathbf{Z}, \Delta, \boldsymbol{\Theta})$, which has two terms. The first, $\Delta\{\mathbf{R} - \mathrm{E}(\mathbf{R} \mid Q_1)\}/\pi(Q_1)$, has mean zero because of the usual Horvitz–Thompson argument, namely complete data ($\Delta = 1$) are being weighted inverse with their selection probabilities $\pi(Q_1)$. The second term has mean zero since $g(\Theta) = \mathrm{E}(R) = \mathrm{E}\{\mathrm{E}(R \mid Q_1)\}$.

This development has one unfortunate catch, namely that, as defined, implementation of (6) is impossible because $\mathcal{T}_1$ and even $\boldsymbol{\chi}$ itself depend on the underlying distributions. There are effectively two ways to implement the procedure:

(a) Use nonparametric regression techniques to estimate $\mathcal{T}_1$ and $\boldsymbol{\chi}$. This gives global asymptotic efficiency but is computationally burdensome and it is unclear if the asymptotics will agree with small sample behavior.

(b) Assume a parametric model for $\mathbf{Z}$ only for the purposes of calculating $\mathcal{T}_1$ and $\boldsymbol{\chi}$. The resulting estimate is (locally) efficient if the parametric model actually holds, and it can be shown that the estimate is consistent even if the assumed parametric model is not correctly specified.

Since for our purposes we are contrasting the various estimators at the normal distribution anyway, we have followed method (b) with $\mathbf{Z}$ assumed to have the multivariate normal distribution. In this case, $\mathcal{T}_1$ and $\boldsymbol{\chi}$ have closed-form expressions that are easily calculated.

We have calculated the asymptotic covariance matrix of the optimal semiparametric estimator when $\mathbf{Z}$ is normally distributed and found that in a wide variety of cases it is essentially the same as that obtained from the Horvitz–Thompson method of moments estimators given in (5). This may not be the case away from the normal distribution, and it is an interesting problem for further study to see if major differences arise with departures from the normal distribution.

3.4 *The AARP Study, Missing Data, and Sampling Weights*

At the first stage of the AARP calibration study, FFQs are mailed to several tens of thousands of members of the AARP randomly selected within certain states. From these FFQs, individuals reporting extreme patterns of food intake are preferentially selected into the calibration study. Suppose that the pattern of intake is quantified by the percent of energy intake contributed by fat (% Calories from Fat). The reported intakes from the FFQ will help to characterize the distribution of a single FFQ report on % Calories from Fat. Suppose we wish to include in the calibration study the following proportions of individuals: 20% having $Q_{i1} \leq 25$, 15% with $25 < Q_{i1} \leq 32.5$, 10% with $32.5 < Q_{i1} \leq 40$, 15% with $40 < Q_{i1} \leq 47.5$, and 40% with $47.5 < Q_{i1}$.

To get some idea of the sampling fractions needed to achieve this, we use the distribution of % Calories from Fat as estimated from the FFQ report in the 1987 National Health Information Survey (NHIS) and in the Women's Health Trial Vanguard Study (Henderson et al., 1990). The estimated mean and variance are 38.25 and 57.76, respectively, and the distribution appears to be reasonably close to normality. We then estimate (using the normality assumption) that, in the AARP study, the selection probabilities should be 1.0, .178, .064, .126, and .962, depending on whether the observed % Calories from Fat lies in 0–25, 25–32.5, 32.5–40, 40–47.5, and >47.5, respectively.

In our numerical work, we will consider the two cases depending on whether the sampling weights are known or estimated. Based on the description in the previous paragraph, to know the weights, we require that the distribution of FFQs is known. This is typically unreasonable in practice, but

in the AARP study, the initial sample size will be so large that, at least at a first level of approximation, the distribution of intakes from FFQs will be effectively known, as will the mean θ_1 and variance θ_3.

In other studies, the initial survey of FFQs will not be so large, and then θ_1, θ_3, and the sampling probabilities must be estimated for all but the normal-theory maximum likelihood estimate. This need not be a bad thing. Robins, Rotnitzky, and Zhao (1994) have shown that such estimation can improve the large-sample properties of Horvitz–Thompson estimators. The obvious nonparametric estimate of the sampling probability in each stratum formed by the initial FFQs is the proportion of individuals in the stratum who are selected into the calibration study.

4. More Individuals or More Food Reports?

When designing the calibration study, there are options regarding the numbers of individuals and how many FRs each case completes. For example, one might obtain just two FRs on many individuals or obtain many FRs but on fewer individuals. Put simply, if one can afford to obtain 4000 FRs, which is the better design option, (a) four FRs on each of 1000 individuals or (b) two FRs on each of 2000 individuals?

Of course, these two designs are not strictly comparable in terms of cost, but they may be nearly so. A design such as (b) incurs more recruitment costs. However, design (a) risks a high dropout rate due to study participants becoming progressively less cooperative; such dropouts not only may bias the analysis but lead to greatly increased costs by attempting to obtain complete records on each individual. Another potential difficulty with observing many FRs on individuals is the possibility of systematic time trends.

As mentioned in the Introduction, the parameters of direct interest in the AARP study are $\rho_{QT} = \theta_4/(\theta_3\theta_6)^{1/2}$, the correlation between intakes from a single FFQ and true usual intake, and the total number N of cancer cases that need to be observed in the main study to achieve 90% power for detecting a plausible and worthwhile effect. Also of interest is the slope β_1. Freedman et al. (1990) give a formula for N.

Whether design option (a) or (b) is preferred may depend on the parameter being estimated as well as on the within-individual error variance in FRs (σ_u^2) relative to the sum of the variance of FFQs about the line (σ_r^2) plus the within-individual error variance in FFQs (σ_ϵ^2). If FRs are relatively precise, then it may be better to select option (b) and maximize the number of individuals in the calibration study. As the FRs become relatively less precise, a switch may occur and it may become more efficient to select design option (a) and take more replicates per person. The switch point may vary according to the parameter being estimated.

We investigate these points first theoretically and then via computer simulation. All calculations use parameters in the model (1)–(2) as estimated via the techniques of Freedman et al. (1991a) for % Calories from Fat as determined by the 1987 NHIS and the Women's Health Trial Vanguard Study (WHTVS), namely $\mu_t = 38.25$, $\sigma_t^2 = 24.45$, $\sigma_\epsilon^2 + \sigma_r^2 = 40.92$, $\sigma_u^2 = 30.36$, $\beta_0 = 5.95$, and $\beta_1 = .83$. These values are consistent with $\text{var}(Q) = 57.76$ mentioned earlier. The WHTVS used food records; if 24-hour recalls are used, σ_u^2 is typically larger, and to incorporate this, we did calculations also in the case that $\sigma_u^2 = 83.35$, a number obtained by an analysis of the CSFII (Continuing Survey of Food Intake by Individuals) data from the U.S. Department of Agriculture.

4.1 *Theoretical Calculations*

We first consider theoretical calculations, which are based on the classical technique of Fisher information theory for the maximum likelihood estimator (Cox and Hinkley, 1981). We assume that, initially, FFQs are obtained on M randomly selected individuals and then FRs are obtained on a calibration subsample of size n. The calculations are standard if this second stage is selected completely at random and if M is infinite so as to essentially completely characterize the distribution of a single FFQ; we will use both assumptions in our theory. Computer simulations will be used to show that the same results apply even when selection into the calibration study depends on the initial FFQ and even if M is finite.

The results of the theoretical calculations are displayed in Figure 1, where we compare design options (a) and (b), described in the previous subsection. We allowed the within-individual error variance in FRs to vary between 0.0 and 150.0 (remember, $\sigma_u^2 \approx 30.36$ for food diaries, while $\sigma_u^2 \approx 83.35$ for 24-hour recalls). As a function of the measurement error variance σ_u^2 in the FRs, this figure compares the ratio of the theoretical (asymptotic) standard deviation for estimates of three parameters of interest: the correlation ρ_{QT}, the slope β_1, and the required number of cancer cases N for $n = 2000$ FRs and $m_2 = 2$ replicates (option b) to $n = 1000$ FRs and $m_2 = 4$ replicates

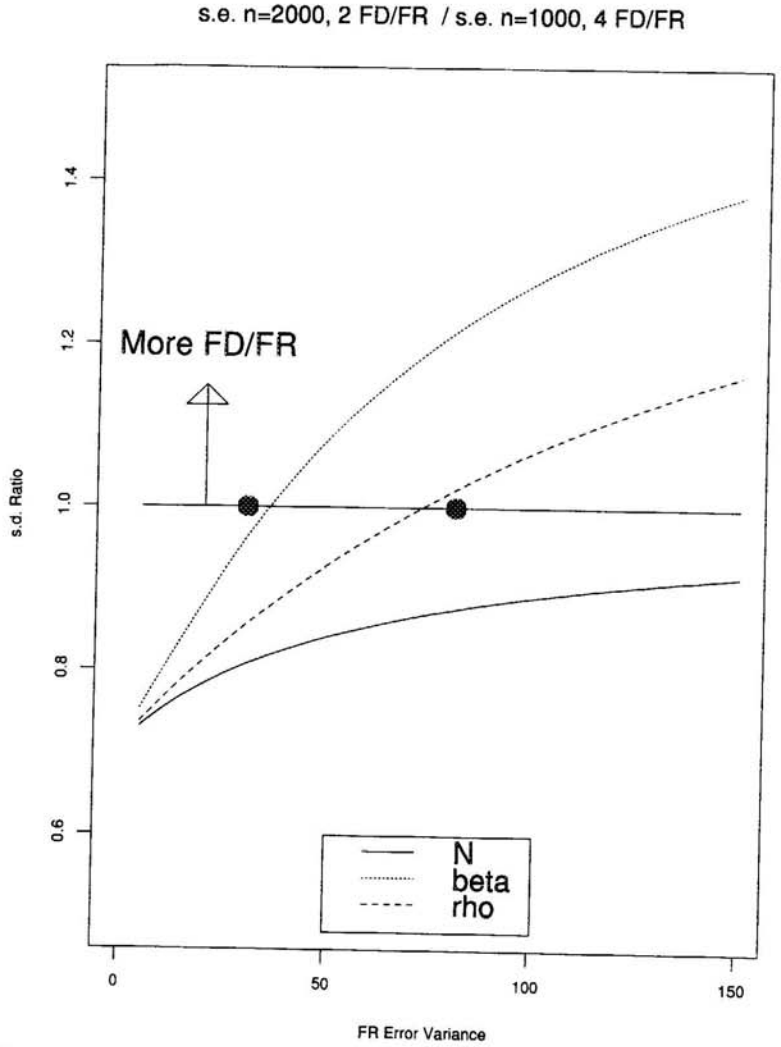

Figure 1. A two-stage calibration study, such as contained within the AARP, with a large number of initial FFQs, under the parameter configurations $\sigma_r^2 = 40.92$, $\mu_t = 38.25$, $\sigma_t^2 = 24.45$, $\beta_0 = 5.95$, and $\beta_1 = .83$. As a function of the measurement error variance σ_u^2 in the FRs, this figure compares the ratio of the asymptotic standard deviation for estimates of ρ_{QT} (rho in the figure), β_1 (beta), and the sample size N (N) for $n = 2000$ FRs and $m_2 = 2$ replicates to $n = 1000$ FRs and $m_2 = 4$ replicates.

(option a). Values of this ratio that are greater than 1.0 indicate that it is better to obtain many FRs on fewer individuals.

The results in Figure 1 are instructive. For our estimate of the within-individual variance of food diaries ($\sigma_u^2 = 30.36$), we see that, for estimating N, β_1, and ρ_{QT}, it is more efficient to obtain fewer food records on many individuals (the ratio is less than 1.0). However, for our estimate of the within-individual variance of 24-hour recalls ($\sigma_u^2 = 83.35$), we see that, especially for β_1, it is more efficient to obtain more recalls on less individuals and, to a lesser extent, the same holds for ρ_{QT}. Interestingly, and for the considered range of the within-individual variance of FRs, for determining the required number of cancer cases N, it is more efficient to obtain only two FRs, both for food diaries and for 24-hour recalls.

The calculations we have done are easily extended in principle to the maximum likelihood estimator under stratified random sampling. Using the theory of estimating equations (Huber, 1967), similar calculations can be performed for the method of moments under either form of sampling.

4.1 *Simulations for Simple Random Selection*

The simulations we have done all agree qualitatively with the theoretical calculations, even when the simulations are applied to stratified sampling (unlike the theory presented here). Numerical results are given in the top half of Tables 1 and 2 for food diaries and in Table 3 for 24-hour recalls. We have listed mean squared errors and standard deviations. There is a small technical problem with listing mean squared errors because Fuller (1987) shows that, in fact, they do not exist theoretically. The problem is that there is a positive probability that the estimated variance of usual intake will equal zero although, with the sample sizes of calibration studies used in the simulations, this chance is so small as to be of no practical concern. Thus, in our particular simulations, this issue was not a problem because we checked the results against a more robust measure of variation, the median absolute deviation from the median, and found no real differences from the results reported here. For smaller calibration studies, this problem of nearly zero estimated variance of usual intake could arise, however. In such cases, the method of moments estimator would be modified as in Section 2.5.1 of Fuller (1987), while the maximum likelihood estimate would probably be best made more stable by Bayesian techniques.

The estimators reported are the normal-theory maximum likelihood estimator and the method of moments estimator, namely solving (5); similar results with respect to design considerations were found for the other distribution-free estimators.

For food records ($\sigma_u^2 = 30.36$), using a larger number of records per individual and fewer individuals is clearly less efficient than using a smaller number of records per individual and more individuals, whether for estimating the slope β_1, the correlation ρ_{QT}, or the required number of cases N. For food recalls ($\sigma_u^2 = 83.35$), we still see that it is more efficient to use fewer rather than

Table 1

Simulation results using food records (FR) and one food frequency questionnaire (FFQ) with $\alpha = 5.95$, $\beta = .83$, $\sigma_r^2 = 16.207$, $\sigma_\epsilon^2 = 24.71$, $\mu_t = 38.253$, $\sigma_t^2 = 24.449$, $\sigma_u^2 = 30.36$. By "Selection on the Basis of FFQ," we mean stratified sampling within ranges of FFQ reported values. The value of n is the number of individuals in the calibration study. The terms ρ_{QT} and $\hat{N}$ refer to the estimate of the correlation between an FFQ and usual intake and the estimated required number of cancer cases, respectively. The method of moments uses the standard parameterization as defined in the text.

		Method of moments				Maximum likelihood			
n	FRs	MSE β	MSE ρ_{QT}	Mean $\hat{N}$	s.d. $\hat{N}$	MSE β	MSE ρ_{QT}	Mean $\hat{N}$	s.d. $\hat{N}$
Selection at Random, $\sigma_u^2 = 30.36$									
200	8	.1014	.0562	2508	721	.0851	.0534	2509	722
400	4	.0840	.0431	2457	501	.0706	.0403	2461	503
800	2	.0707	.0358	2393	390	.0648	.0344	2396	390
500	8	.0646	.0359	2441	433	.0541	.0342	2442	433
1000	4	.0497	.0271	2397	302	.0431	.0260	2398	301
2000	2	.0489	.0239	2389	240	.0443	.0227	2391	240
800	8	.0512	.0279	2385	313	.0408	.0261	2386	312
1600	4	.0407	.0221	2390	240	.0351	.0211	2391	241
3200	2	.0368	.0181	2380	188	.0337	.0174	2380	188
Selection on the Basis of FFQ, $\sigma_u^2 = 30.36$									
200	8	.1192	.0581	2484	649	.0785	.0373	2415	403
400	4	.0975	.0443	2430	475	.0656	.0312	2395	322
800	2	.0992	.0430	2408	393	.0673	.0282	2385	249
500	8	.0744	.0357	2417	389	.0482	.0232	2387	250
1000	4	.0629	.0295	2398	302	.0416	.0193	2385	193
2000	2	.0585	.0255	2395	239	.0389	.0168	2384	157
800	8	.0590	.0288	2377	288	.0373	.0184	2380	193
1600	4	.0496	.0230	2389	221	.0317	.0148	2377	147
3200	2	.0481	.0209	2384	192	.0311	.0133	2375	125

Table 2

Simulation results using food records (FR) and two food frequency questionnaires (FFQ) with $\alpha = 5.95$, $\beta = .83$, $\sigma_\tau^2 = 16.207$, $\sigma_\epsilon^2 = 24.71$, $\mu_t = 38.253$, $\sigma_t^2 = 24.449$, $\sigma_u^2 = 30.36$. *See Table 1 for definition of terms.*

		Method of moments				Maximum likelihood			
n	FRs	MSE β	MSE ρ_{QT}	Mean $\hat{N}$	s.d. $\hat{N}$	MSE β	MSE ρ_{QT}	Mean $\hat{N}$	s.d. $\hat{N}$
Selection at Random, $\sigma_u^2 = 30.36$									
200	8	.0921	.0498	2499	624	.0733	.0451	2496	618
400	4	.0693	.0373	2436	448	.0610	.0349	2434	444
800	2	.0651	.0304	2403	333	.0615	.0290	2405	332
500	8	.0544	.0299	2391	355	.0444	.0269	2390	349
1000	4	.0442	.0230	2396	260	.0373	.0210	2395	258
2000	2	.0405	.0198	2393	209	.0373	.0186	2393	206
800	8	.0421	.0236	2391	273	.0340	.0214	2388	270
1600	4	.0362	.0193	2377	208	.0305	.0174	2375	203
3200	2	.0336	.0161	2380	160	.0309	.0151	2380	167
Selection on the Basis of FFQ, $\sigma_u^2 = 30.36$									
200	8	.1109	.0537	2471	659	.0700	.0339	2413	392
400	4	.0919	.0419	2451	478	.0583	.0263	2396	281
800	2	.0948	.0392	2413	372	.0620	.0251	2386	225
500	8	.0682	.0331	2410	393	.0431	.0209	2395	239
1000	4	.0565	.0265	2387	289	.0357	.0167	2375	181
2000	2	.0572	.0244	2387	236	.0367	.0151	2376	144
800	8	.0549	.0271	2421	309	.0337	.0170	2389	193
1600	4	.0442	.0211	2373	226	.0293	.0136	2376	146
3200	2	.0453	.0200	2386	192	.0307	.0123	2373	112

Table 3

Simulation results using 24-hour recalls (24-FR) with $\alpha = 5.95$, $\beta = .83$, $\sigma_\tau^2 = 16.207$, $\sigma_\epsilon^2 = 24.71$, $\mu_t = 38.253$, $\sigma_t^2 = 24.449$, $\sigma_u^2 = 83.35$. *By "Selection on the Basis of FFQ," we mean stratified sampling within ranges of FFQ reported values. The value of n is the number of individuals in the calibration study. The terms ρ_{QT} and $\hat{N}$ refer to the estimate of the correlation between an FFQ and usual intake and the estimated required number of cancer cases, respectively.*

		Method of moments				Maximum likelihood			
n	FRs	MSE β	MSE ρ_{QT}	Mean $\hat{N}$	s.d. $\hat{N}$	MSE β	MSE ρ_{QT}	Mean $\hat{N}$	s.d. $\hat{N}$
Selection at Random, $\sigma_u^2 = 83.35$									
500	8	.0759	.0408	2404	473	.0679	.0394	2406	473
1000	4	.0738	.0363	2403	379	.0689	.0349	2407	378
2000	2	.0832	.0356	2387	330	.0817	.0354	2391	330
Selection on the Basis of FFQ, $\sigma_u^2 = 83.35$									
500	8	.1022	.0443	2419	443	.0632	.0288	2396	291
1000	4	.1051	.0424	2402	358	.0662	.0272	2383	231
2000	2	.1392	.0517	2404	340	.0814	.0302	2372	206
Selection at Random, $\sigma_u^2 = 83.35$									
500	8	.0657	.0342	2397	396	.0567	.0312	2395	390
1000	4	.0650	.0303	2401	326	.0628	.0296	2404	326
2000	2	.0816	.0324	2398	287	.0800	.0318	2390	286
Selection on the Basis of FFQ, $\sigma_u^2 = 83.35$									
500	8	.0917	.0413	2410	451	.0594	.0263	2383	273
1000	4	.0912	.0386	2420	374	.0588	.0240	2395	227
2000	2	.1247	.0464	2420	359	.0801	.0284	2393	194

more records per individual for estimating the required number of cases, but $m = 4$ records per individual appears to be somewhat more efficient than $m = 2$ records per individual for estimating the slope β_1 and the correlation ρ_{QT}.

4.3 *Effects of Stratified Sampling*

In the AARP calibration study, stratified sampling appears attractive so as to parallel the stratified nature of the main study. Here we study the statistical issue of stratified versus completely random sampling into the calibration substudy (Tables 1–3).

The results are striking. Both asymptotic theory (not reported here) and simulations indicate that, for the distribution-free methods, stratification causes a decrease in efficiency of estimation, particularly for the slope β_1 and in some cases for the correlation ρ_{QT}. There is also some decrease in efficiency for estimating the required number of cancer cases N, although the effect is not large.

Exactly the opposite results obtain for the normal-theory maximum likelihood estimator. Here we see that there is not much effect due to the design for estimating the slope β_1, but now the stratified design leads to noticeably smaller variability in the correlation ρ_{QT} and the required number of cancer cases N.

4.4 *Number of FFQs*

Tables 1–3 are simulations based on an infinite number (M) of initial FFQs. Obviously, in practice M will be finite, so we ran simulations with $M = 15{,}000$ initial FFQs, which is the target for the AARP study. The results are reported in Table 4 and should be compared to Table 1. There are essentially no differences between the tabulated value for $M = \infty$ and $M = 15{,}000$.

5. Parametric or Semiparametric Analysis?

As stated previously, under simple random sampling into the calibration study, our experience in this and other problems has been that distribution-free estimates and the normal-theory maximum likelihood estimate behave similarly. Tables 1–3 report results for random selection for the maximum likelihood estimator and the method of moments estimator (5), and while there are some differences, they are typically fairly minor, as expected. The semiparametric efficient estimator (6) is equivalent to the maximum likelihood estimator in this case.

It is when selection into the calibration study depends on the initial response that we see major differences (Tables 1–3). While we report results only for the method of moments estimator (5), the semiparametric efficient distribution-free method (6) gave similar results. Both are vastly inferior to the normal-theory maximum likelihood estimator. For estimating the slope β_1, the correlation ρ_{QT}, or the required number of cancer cases N, the maximum likelihood estimator has less than 50% of the variance of the semiparametric methods. The inefficiency of the semiparametric estim-

Table 4

Simulation results using two food records (FR, $\sigma_u^2 = 30.36$) and two 24-hour recalls (24-FR, $\sigma_u^2 = 83.35$) with $\alpha = 5.95$, $\beta = .83$, $\sigma_r^2 = 16.207$, $\sigma_\epsilon^2 = 24.71$, $\mu_t = 38.253$, $\sigma_t^2 = 24.449$. Here we use stratified sampling within ranges of FFQ reported values, where the initial survey consists of 15,000 FFQs. The value of n is the number of individuals in the calibration study. The terms ρ_{QT} and $\hat{N}$ refer to the estimate of the correlation between an FFQ and usual intake and the estimated required number of cancer cases, respectively.

		Method of moments				Maximum likelihood			
n	FRs	MSE β	MSE ρ_{QT}	Mean $\hat{N}$	s.d. $\hat{N}$	MSE β	MSE ρ_{QT}	Mean $\hat{N}$	s.d. $\hat{N}$
		Single FFQ, Two FRs							
		Selection on the Basis of 15,000 Initial FFQs, $\sigma_u^2 = 30.36$							
500	8	.0744	.0368	2393	384	.0492	.0236	2381	250
1000	4	.0600	.0287	2390	299	.0405	.0191	2377	192
2000	2	.0602	.0261	2407	244	.0384	.0166	2381	158
		Single FFQ, Two 24-FRs							
		Selection on the Basis of 15,000 Initial FFQs, $\sigma_u^2 = 83.35$							
500	8	.0966	.0443	2426	453	.0631	.0292	2391	292
1000	4	.0979	.0415	2444	394	.0659	.0271	2392	245
2000	2	.1288	.0484	2416	359	.0851	.0304	2389	207

ator is not due to the implementation described in Section 3.3 but instead is due to the nonparametric aspect of the procedure.

6. Discussion

A major point of our paper is that calibration studies may not always be designed simply to provide data to adjust relative risks in a larger study, although we agree that such adjustment is indeed an important aspect. For example, calibration studies should be used in the development of new measurement instruments to test whether the new measurement provides improvement over currently used methods. In this context, the correlation between the instrument and the true measurement (cf., ρ_{QT}) would be of primary interest, and measures of bias (cf., β_0 and β_1) would also be important. As in the AARP study, calibration substudies may also be planned as part of the design phase in which design assumptions are checked. In this case, a central question is whether the designed sample size of the main study is justified. Our paper therefore addresses the design of calibration studies from a wider perspective than heretofore.

With regard to the question of the number of food reports per individual in the calibration study, our results are summarized in Table 5. They suggest that the conclusions of Kaaks et al. (1995) and Stram et al. (1995) that fewer repeat measurements on more individuals provides greater efficiency is not completely general. Indeed, this was already demonstrated by Rosner and Willett (1988), who showed that, for estimating the correlation between an FFQ and usual intake, the size of the measurement error in the FFQ and the FRs determine the optimal strategy. We have demonstrated as well that the optimal balance of repeats and individuals depends on the primary aim of the calibration study as well as on the within-individual variation of the repeated measurements. We should note, however, that for the particular parameters of our simulations, there were no cases where the choice of two repeats per individual was much worse than four repeats on half the number of individuals, indicating that in our case the amount of within-individual variation, even in 24-hour recalls, was not enough to depart from the policy of maximizing the number of individuals. From Figure 1, though, it is clearly quite possible that such a policy could be seriously in error in other circumstances.

It is worth emphasizing that, as described in Section 1.2, Kaaks et al. (1995) and Stram et al. (1995) have shown that, for estimating the attenuation in relative risk due to using FFQs, i.e., the slope in the regression of usual intake on the FFQ, the optimal choice is to obtain only a single FR in the calibration study. Independent of whether the calibration study is based on simple random or stratified sampling, the natural (and normal-theory maximum likelihood estimate) of this slope is the ordinary least squares regression slope when regressing the mean of the within-person FRs on the FFQ. If one has a choice between a calibration study of size n with k FRs per individual or a calibration study of size nk with one FR per individual, then, independent of the sampling design (simple random or stratified), the ratio of the variance of the attenuation for the first strategy relative to the second is

$$\frac{k(1-\rho_{QT}^2+\omega/k)}{1-\rho_{QT}^2+\omega},$$

Table 5

For various problems, the minimum number of FRs required in a calibration study (Min number) and the optimal number (Optimal number)

Problem	Min number	Optimal number
Correcting relative risks using regression calibration	1	1
Estimating required number of cases to detect effect at a given power	1 or 2, depending on method used	Same as Min number
Estimating correlation ρ_{QT}	2	No uniform answer
Estimating slope β_1	2	No uniform answer

where $\omega = \sigma_u^2/\sigma_t^2$ (see Kaaks et al., 1995). For the parameter configurations used in our simulations, the ratio is 1.35 for $k = 2$ versus $k = 1$ FRs per individual when using food records and 1.17 when using 24-hour recalls.

The results regarding the advisability of stratified sampling into calibration studies do not provide a clear answer because gains in efficiency are made under one analysis strategy (maximum likelihood) and losses are made under the other strategy (method of moments). As we have emphasized, the likelihood approach is valid only if the parametric structure is correctly specified. Likelihood methods require statistical models for the distribution of the true variate T. There has traditionally been considerable concern in the measurement error literature about the robustness of estimation and inferences based on parametric models for unobservable variates. Fuller (1987, p. 263) discusses this issue briefly in the classic nonlinear regression problem and basically concludes that the results of parametric modeling "may depend heavily on the (assumed) form of the (T) distribution." In probit regression, Carroll et al. (1984) report that, if one assumes T is normally distributed and it really follows a chi-squared distribution with one degree of freedom, then the effect on the likelihood estimate is markedly negative. Similar results are reported by Schafer (1987). Essentially, all research workers in the measurement error field come to a common conclusion: likelihood methods can be of considerable value, but the possible nonrobustness of inference due to model misspecification is a vexing and difficult problem.

The issue of model robustness is hardly limited to measurement error modeling. Indeed, it pervades statistics and has led to the rise of a variety of semiparametric and nonparametric techniques. There is simply no agreement in the statistical literature as to whether semi/nonparametric or parametric modeling is more appropriate. Many researchers strongly believe that one should make as few model assumptions as possible. The argument here is that any extra efficiency gained by parametric modeling is more than offset by the need to perform careful and often time-consuming sensitivity analyses. Other researchers believe that appropriate statistical analysis requires one to do one's best to model every feature of the data, arguing in our context that it makes little sense to needlessly double the variance of parameter estimates.

The obvious question is whether the maximum likelihood estimate is actually sensitive in this context to model misspecification. We have run simulations with T having a scaled and translated negative exponential distribution and found that the normal-theory maximum likelihood estimate of β_1 is badly biased downwards, and this translates into a bias in the estimate of ρ_{QT}. For instance, for the parameters in Table 1, $\beta_1 = .83$ and $\rho_{QT} = .54$, while in the simulations the averages are .62 and .49, respectively. While these biases are considerable, we note that, based on simulations, the 5%-level Anderson–Darling test for normality has power over 80% for detecting the nonnormality caused by the nonnormal distribution of T for as few as 2000 FFQs, and for larger sample sizes such as in the AARP study, the power is nearly 100%. The point here is that, while a misspecified likelihood analysis leads to badly biased estimates, in practice, it is not impossible to detect the model misspecification, even with the large amounts of measurement error inherent in nutritional intake data.

A practical question is whether one can ever reasonably assume normality. With a stratified design, of course, the observed FRs will not be normally distributed anyway, and so distributional modeling is easiest for the FFQs. It is often the case that nutrition data are transformed directly to normality (Nusser et al., 1996, give one such approach), and the analysis is then done on the transformed scale. If one is willing to assume that when transformed FFQs are normally distributed so too are their (transformed) component parts T and ϵ as well as the FRs, then the modeling issue is solved.

A referee has also brought up the valuable point that one way to check the distributional model in stratified samples, albeit indirectly, is to evaluate the linearity and normality of the regression of FRs on FFQs.

For % Calories from Fat, the nutrient intakes from many data sets appear reasonably normally distributed. In Table 6, we review the evidence of five studies with various instruments, noting that, for the eight situations surveyed, six are reasonably normally distributed (and pass the Anderson–Darling test with level $>.05$), one exhibits light-tailedness (Nurses Health Study, 4-day diaries), and another appears to be heavy-tailed (NHANES, 24-hour recalls). The latter is the only situation where one might expect that a parametric analysis assuming normality might be badly biased.

Since, as we have argued above, % Calories from Fat often does appear to follow a normal distribution, our results indicate that in our case it would make sense to adopt a maximum likelihood approach, and consequently, stratified sampling would appear beneficial. In general, however, we are not foolhardy enough to recommend one or the other approach. The important point is that

Table 6

Tests for normality for the variable % Calories from Fat for various nutrition data sets. The 5% significance level for the A–D (Anderson–Darling) test is .78. Definition of acronyms: WISH (Women's Interview Survey of Health), CSFII (Continuing Survey of Food Intake by Individuals), NHANES (National Health and Nutrition Examination Survey), NHS (Nurses' Health Study), WHTVS (Women's Health Trial Vanguard Study).

Study	Instrument	No. instruments per participant	Sample size	Skewness	Kurtosis	A–D Test
WISH	FFQ	1	271	−.12	3.24	.68
WISH	24-hour recall	6	271	−.31	3.28	.49
CSFII	24-hour recall	3	1705	.01	3.42	.71
NHANES	24-hour recall	1	3145	.08	3.65	1.61
NHS	FFQ	1	168	.13	3.60	.29
NHS	4-day diary	4	168	−.29	2.38	.91
WHTVS	FFQ	3	86	.16	2.58	.42
WHTVS	4-day diary	2	86	−.55	3.35	.47

we have identified a practical problem in which there is a surprisingly large difference between parametric and semiparametric modeling.

Acknowledgements

Carroll's research was supported by a grant from the National Cancer Institute (CA-57030). It was completed during a visit to Sonderforschungsbereich 373, Institut für Statistik und Ökonometrie, Humboldt Universität zu Berlin, under the auspices of a senior Alexander von Humboldt Foundation Research Award.

Résumé

Nous étudions, en nous appuyant sur un exemple en épidémiologie de la nutrition, quelques aspects d'analyse et de d'élaboration de mesures d'erreurs dans des modèles linéaires avec des données auxiliaires manquantes. Le contexte étudié est celui d'un grand échantillon initial dans lequel la réponse (un questionnaire de fréquence alimentaire, FFQ) est observée ainsi qu'une étude de calibration moins importante dans laquelle on dispose de répétitions de l'erreur de prédiction (enregistrements alimentaires ou rappels sur consommation, FR). La différence entre notre analyse et la plupart des modèles de mesure d'erreurs de la littérature est que, dans notre étude, la sélection dans l'étude de calibration peut dépendre de la valeur de la réponse. Un exposé raisonné de co type d'étude est présenté. Deux problèmes majours sont étudiés. Dans la conception d'une étude de calibration nous avons la possibilité d'avoir soit des échantillons de grande taille et peu de répétitions ou bien des échantillons de petite taille et plus de répétitions. Il est montré, de façon un peu surprenante, qu'aucune stratégle n'est uniformément meilleure dans des cas pratiques. Les réponses dépendent de l'instrument utilisé (rappels ou enregistrements) et des paramètres d'intérêt. Le second point étudié est un problème d'analyse. Dans des modèles linéaires usuels sans données manquantes les estimations obtenues par la méthode des moments et ceux de la théorie normale du maximum de vraisemblance sont approximativement équivalents, avec une utilisation plus fréquente de la première méthodes en raison d'un calcul facile et explicite. Les deux estimations sont valides sans besoin d'hypothèses distributionnelles. Par contre, dans le cas de données manquantes, seule la méthode des moments reste applicable sans nécessiter de poser des hypothèses distributionnelles alors que l'estimation du maximum de vraisemblance a une précision au moins 50% plus importante dans des situations pratiques lorsque l'on peut admettre la normalité. Les conséquences pour les études de calibration de dispositifs d'enquêtes en nutrition sont discutées.

References

Bingham, S. A. (1991). Limitations of the various methods for collecting dietary intake data. *Annals of Nutritional Metabolism* **35,** 117–127.

Carroll, R. J., Spiegelman, C., Lan, K. K., Bailey, K. T., and Abbott, R. D. (1984). On errors-in-variables for binary regression models. *Biometrika* **71,** 19–26.

Carroll, R. J., Freedman, L., and Hartman, A. (1995). The use of semiquantitative food frequency questionnaires to estimate the distribution of usual intake. *American Journal of Epidemiology* **143,** 392–404.

Carroll, R. J., Freedman, L. S., Kipnis, V., and Li, L. (1998). A new class of measurement error models, with applications to dietary data. *Canadian Journal of Statistics,* in press.

Cox, D. R. and Hinkley, D. V. (1981). *Theoretical Statistics.* London: Academic Press.

Freedman, L., Schatzkin, A., and Wax, Y. (1990). The impact of dietary measurement error on planning the sample size required in a cohort study. *American Journal of Epidemiology* **132,** 1185–1195.

Freedman, L. S., Carroll, R. J., and Wax, Y. (1991). Estimating the relationship between dietary intake obtained from a food frequency questionnaire and true average intake. *American Journal of Epidemiology* **134,** 510–520.

Freedman, L. S., Schatzkin, A., and Wax, Y. (1991). Re: The impact of dietary measurement error on planning sample size required in a cohort study. The authors reply. *American Journal of Epidemiology* **134,** 1472–1473.

Freudenheim, J. L. and Marshall, J. R. (1988). The problem of profound mismeasurement and the power of epidemiologic studies of diet and cancer. *Nutrition and Cancer* **11,** 243–250.

Fuller, W. A. (1987). *Measurement Error Models.* New York: John Wiley and Sons.

Hebert, J. R. and Miller, D. R. (1988). Methodologic considerations for investigating the diet–cancer link. *American Journal of Clinical Nutrition* **47,** 1068–1077.

Heitmann, B. L. and Lissner, L. (1995). Dietary underreporting by obese individuals: Is it specific or non-specific? *British Medical Journal* **311,** 986–989.

Henderson, M. M., Kushi, L. H., Thompson, D. J., et al. (1990). Feasibility of a randomized trial of a low-fat diet for the prevention of breast cancer: Dietary compliance in the Women's Health Trial Vanguard Study. *Preventive Medicine* **19,** 115–133.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47,** 663–685.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the 5th Berkeley Symposium* **1,** 221–233.

Kaaks, R., Riboli, E., and van Staveren, W. (1995). Sample size requirements for calibration studies of dietary intake measurements in prospective cohort investigations. *American Journal of Epidemiology* **142,** 557–565.

Kent, J. T. (1982). Robust properties of likelihood ratio tests. *Biometrika* **69,** 19–27.

Nusser, S. M., Carriquiry, A. L., Dodd, K. W., and Fuller, W. A. (1996). A semiparametric transformation approach to estimating usual intake distributions. *Journal of the American Statistical Association* **91,** 1040–1046.

Plummer, M. and Clayton, D. (1993). Measurement error in dietary assessment: An investigation using covariance structure models, part II. *Statistics in Medicine* **12,** 937–948.

Prentice, R. L. (1996). Dietary fat and breast cancer: Measurement error and results from analytic epidemiology. *Journal of the National Cancer Institute* **88,** 1738–1747.

Prentice, R. L., Kakar, F., Hursting, S., Sheppard, L., Klein, R., and Kushi, L. H. (1988). Aspects of the rationale for the Women's Health Trial. *Journal of the National Cancer Institute* **80,** 802–814.

Rosner, B. and Willett, W. C. (1988). Interval estimates for correlation coefficients corrected for within-person variation: Implications for study design and hypothesis testing. *American Journal of Epidemiology* **127,** 377–386.

Rosner, B. A., Willett, W. C., and Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine* **8,** 1051–1070.

Rotnitzky, A. and Robins, J. M. (1995). Semiparametric estimation of models for means and covariances in the presence of missing data. *Scandanavian Journal of Statistics* **22,** 323–333.

Schafer, D. (1987). Covariate measurement error in generalized linear models. *Biometrika* **74,** 385–391.

Stram, D. O., Longnecker, M. P., and Shames, L. (1995). Cost-efficient design of a diet validation study. *American Journal of Epidemiology* **142,** 353–362.

Willett, W. C., Sampson, L., and Stampfer, M. J. (1985). Reproducibility and validity of a semiquantitative food frequency questionnaire. *American Journal of Epidemiology* **122,** 51–65.

Received May 1996; *revised March* 1997; *accepted July* 1997.

APPENDIX

A1. Estimating Function for the Normal-Theory Maximum Likelihood

We first write the estimating function and its derivatives for the case in which there are no missing data. Recall that $\boldsymbol{\Sigma}(\boldsymbol{\Theta})$ is the covariance matrix of all the data (see (3)). Let the partial derivatives of this matrix with respect to an arbitrary θ_j be given by

$$\mathbf{P}_j = \frac{\partial}{\partial \theta_j}\boldsymbol{\Sigma}(\boldsymbol{\Theta}),$$

where the derivative is componentwise. Then, using matrix derivatives, the estimating function for the maximum likelihood estimator when computed on all the data $\mathbf{Z}$ can be shown to equal (the numerical ordering appears slightly odd but is convenient later)

$$\boldsymbol{\Psi}_{\mathrm{ml}}(\mathbf{Z},\boldsymbol{\Theta}) = \begin{bmatrix} \boldsymbol{\Psi}_{\mathrm{ml},1}(\mathbf{Z},\boldsymbol{\Theta}) \\ \boldsymbol{\Psi}_{\mathrm{ml},2}(\mathbf{Z},\boldsymbol{\Theta}) \end{bmatrix} = \{\ell_1(\mathbf{Z},\boldsymbol{\Theta}), \ell_3(\mathbf{Z},\boldsymbol{\Theta}), \ell_2(\mathbf{Z},\boldsymbol{\Theta}), \ell_4(\mathbf{Z},\boldsymbol{\Theta}), \ldots\}^t,$$

where

$$\ell_1(\mathbf{Z},\boldsymbol{\Theta}) = \begin{pmatrix} \mathbf{e}_{m_1} \\ 0\cdot\mathbf{e}_{m_2} \end{pmatrix}^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Theta}) \left\{ \mathbf{Z} - \begin{pmatrix} \theta_1\mathbf{e}_{m_1} \\ \theta_2\mathbf{e}_{m_2} \end{pmatrix} \right\},$$

$$\ell_2(\mathbf{Z},\boldsymbol{\Theta}) = \begin{pmatrix} 0\cdot\mathbf{e}_{m_1} \\ \mathbf{e}_{m_2} \end{pmatrix}^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Theta}) \left\{ \mathbf{Z} - \begin{pmatrix} \theta_1\mathbf{e}_{m_1} \\ \theta_2\mathbf{e}_{m_2} \end{pmatrix} \right\},$$

and, for $j \geq 3$,

$$\begin{aligned} \ell_j(\mathbf{Z},\boldsymbol{\Theta}) = & -(1/2)\mathrm{trace}\left\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Theta})\mathbf{P}_j\right\} \\ & + (1/2)\left\{ \mathbf{Z} - \begin{pmatrix} \theta_1\mathbf{e}_{m_1} \\ \theta_2\mathbf{e}_{m_2} \end{pmatrix} \right\}^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Theta})\mathbf{P}_j\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Theta}) \left\{ \mathbf{Z} - \begin{pmatrix} \theta_1\mathbf{e}_{m_1} \\ \theta_2\mathbf{e}_{m_2} \end{pmatrix} \right\}. \end{aligned}$$

Again using matrix derivatives, the Hessian of the m_1 FFQs and the m_2 FRs can be computed explicitly as follows. Let

$$\mathbf{A}(\theta_1,\theta_2) = \mathbf{Z}\mathbf{Z}^t - \mathbf{Z}\begin{pmatrix} \theta_1\mathbf{e}_{m_1} \\ \theta_2\mathbf{e}_{m_2} \end{pmatrix}^t - \begin{pmatrix} \theta_1\mathbf{e}_{m_1} \\ \theta_2\mathbf{e}_{m_2} \end{pmatrix}\mathbf{Z}^t + \begin{pmatrix} \theta_1\mathbf{e}_{m_1} \\ \theta_2\mathbf{e}_{m_2} \end{pmatrix}\begin{pmatrix} \theta_1\mathbf{e}_{m_1} \\ \theta_2\mathbf{e}_{m_2} \end{pmatrix}^t.$$

The Hessian of $(\ell_1, \ell_2, \ldots, \ell_7)$ is a 7×7 matrix with elements $h_{jk}(\boldsymbol{\Theta})$, where

$$h_{11} = -\begin{pmatrix} \mathbf{e}_{m_1} \\ 0\cdot\mathbf{e}_{m_2} \end{pmatrix}^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Theta}) \begin{pmatrix} \mathbf{e}_{m_1} \\ 0\cdot\mathbf{e}_{m_2} \end{pmatrix},$$

$$h_{12} = -\begin{pmatrix} \mathbf{e}_{m_1} \\ 0\cdot\mathbf{e}_{m_2} \end{pmatrix}^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Theta}) \begin{pmatrix} 0\cdot\mathbf{e}_{m_1} \\ \mathbf{e}_{m_2} \end{pmatrix},$$

$$h_{22} = -\begin{pmatrix} 0\cdot\mathbf{e}_{m_1} \\ \mathbf{e}_{m_2} \end{pmatrix}^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Theta}) \begin{pmatrix} 0\cdot\mathbf{e}_{m_1} \\ \mathbf{e}_{m_2} \end{pmatrix},$$

and, for $j, k \geq 3$,

$$h_{1j} = -\begin{pmatrix} \mathbf{e}_{m_1} \\ 0\cdot\mathbf{e}_{m_2} \end{pmatrix}^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Theta})\mathbf{P}_j\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Theta}) \left\{ \mathbf{Z} - \begin{pmatrix} \theta_1\mathbf{e}_{m_1} \\ \theta_2\mathbf{e}_{m_2} \end{pmatrix} \right\},$$

$$h_{2j} = -\begin{pmatrix} 0\cdot\mathbf{e}_{m_1} \\ \mathbf{e}_{m_2} \end{pmatrix}^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Theta})\mathbf{P}_j\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Theta}) \left\{ \mathbf{Z} - \begin{pmatrix} \theta_1\mathbf{e}_{m_1} \\ \theta_2\mathbf{e}_{m_2} \end{pmatrix} \right\},$$

$$\begin{aligned} h_{jk} = & (1/2)\mathrm{trace}\left\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Theta})\mathbf{P}_k\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Theta})\mathbf{P}_j\right\} \\ & - (1/2)\mathrm{trace}\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Theta})\left\{\mathbf{P}_k\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Theta})\mathbf{P}_j + \mathbf{P}_j\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Theta})\mathbf{P}_k\right\}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Theta})\mathbf{A}(\theta_1,\theta_2)\right]. \end{aligned}$$

Now we consider the possibility of missing data. From Little and Rubin (1987), the maximum likelihood estimator does not take into account the selection probabilities and hence solves

$$0 = \sum_{i=1}^{M}\left(\Delta_i\begin{bmatrix} \boldsymbol{\Psi}_{\mathrm{ml},1}(\mathbf{Z}_i,\boldsymbol{\Theta}) \\ \boldsymbol{\Psi}_{\mathrm{ml},2}(\mathbf{Z}_i,\boldsymbol{\Theta}) \end{bmatrix} + (1-\Delta_i)\begin{bmatrix} \boldsymbol{\Psi}_{\mathrm{ml},3}(Q_{i1},\boldsymbol{\Theta}) \\ \mathbf{0} \end{bmatrix}\right),$$

where

$$\Psi_{\mathrm{ml},3}(Q_{i1}, \boldsymbol{\Theta}) = \begin{bmatrix} (Q_{i1} - \theta_1)/\theta_3 \\ (2\theta_3^2)^{-1}\left\{(Q_{i1} - \theta_1)^2 - \theta_3\right\} \end{bmatrix}.$$

A.2. Inconsistency of the MLE for Nonnormal Distributions

Showing this fact algebraically is a somewhat unpleasant task in general, but a simple special case illustrates the main idea. Suppose that the mean and variance of Q are known; this never happens exactly, but in the AARP study $n > 300{,}000$, and so for all realistic purposes, the mean and variance of Q really is known. For simplicity, suppose that σ_u^2 is known and that $m_2 = 1$, i.e., there is only one FR. Recalling that $\theta_1 = \mathrm{E}(Q) = \beta_0 + \beta_1\mu_t$, $\theta_2 = \mathrm{E}(F) = \mu_t$, $\theta_3 = V(Q) = \beta_1^t\sigma_t^2 + \sigma_r^2$, $\theta_4 = \beta_1\sigma_t^2$, and $\theta_5 = \sigma_t^2 + \sigma_u^2$, the unknown parameters for a normal-theory likelihood analysis are $(\theta_2, \theta_4, \theta_5)$. By detailed algebra, it may be shown that the normal-theory maximum likelihood estimator of θ_2 must satisfy

$$0 = \sum_{i=1}^{n} \Delta_i \left\{F_{i1} - \theta_2 - (\theta_4/\theta_3)(Q_{i1} - \theta_1)\right\}.$$

By the usual theory of estimating equations, if all parameters can be estimated consistently, then

$$0 = \mathrm{E}\left[\Delta\left\{F - \theta_2 - (\theta_4/\theta_3)(Q - \theta_1)\right\}\right].$$

Remembering that $\mathrm{E}(\Delta \mid F, Q) = \mathrm{E}(\Delta \mid Q) = \pi(Q)$ because the data are missing at random, we see that consistency requires that

$$0 = \mathrm{E}\left[\pi(Q)\left\{\mathrm{E}(F \mid Q) - \theta_2 - (\theta_4/\theta_3)(Q - \theta_1)\right\}\right]. \tag{7}$$

Note that (7) holds if $\pi(Q) \equiv \pi$, a constant. In general, however, because the function $\pi(\cdot)$ is arbitrary, for (7) to hold we require that the regression of F on Q be linear. This reflects the distributional assumption of normality and need not hold otherwise.